

PYTHON UYGULAMALI İSTATİKSEL VERİ BİLİMİ VE ANALİZİ

Yazar

Dr. Ahmet SEL

© Copyright 2021

Bu kitabın, basım, yayın ve satış hakları Akademisyen Kitabevi A.Ş.'ne aittir. Anılan kuruluşun izni alınmadan kitabın tümü ya da bölümleri mekanik, elektronik, fotokopi, manyetik kâğıt ve/veya başka yöntemlerle çoğaltılamaz, basılamaz, dağıtılamaz. Tablo, şekil ve grafikler izin alınmadan, ticari amaçlı kullanılamaz. Bu kitap T.C. Kültür Bakanlığı bandrolü ile satılmaktadır.

ISBN

978-625-7409-60-5

Kitap Adı

Python Uygulamalı İstatistiksel Veri Bilimi ve Analizi

Yazarlar

Ahmet SEL

ORCID iD: 0000-0003-1914-5878

Yayın Koordinatörü

Yasin DİLMEN

Sayfa ve Kapak Tasarımı

Akademisyen Dizgi Ünitesi

Yayıncı Sertifika No

47518

Baskı ve Cilt

Vadi Matbaacılık

Bisac Code

SOC027000

DOI

10.37609/akya.286

GENEL DAĞITIM

Akademisyen Kitabevi A.Ş.

Halk Sokak 5 / A

Yenişehir / Ankara

Tel: 0312 431 16 33

siparis@akademisyen.com

www.akademisyen.com

ÖNSÖZ

Bilimsel gelişmelerin hızlı yaşandığı günümüzde giderek daha fazla veri toplanır ve depolanır hale gelmiştir. İlk başta veriler; sayı, metin, resim, video ve daha birçok kaynaktan elde edilen anlamsız bir yığın olarak görülebilir. Büyük veri, değişik kaynaklardan toparlanan ham verilerin anlamlı ve işlenebilir biçime dönüştürülmüş biçimine denir. Veri Bilimi; verilerin işlenmesi, analiz edilmesi ve yorumlanması adımlarını kullanarak faydalı bilgilerin elde edilmesini sağlar. Verilerin analiz edilmesinde istatistiksel yöntemlerin kullanılması verilerin yorumlanmasında hatalı sonuçların elde edilmesini önler. Çalışmada bu amaçla; veri temizleme, veri hazırlama, veri görselleştirme ve veri analizi dahil olmak üzere istatistiksel yöntemler kullanılmıştır.

Bu kitapta sunulan pratik kavramlar karar verme süreçlerinde bilgiyi geliştirmek ve yorumlamak için sosyoloji, pazarlama, işletme, kalite kontrol, eğitim, ekonomi, tıp ve mühendislik gibi veri analizi ve sentezinin gerekli olduğu birçok bilim dalında faydalı olacağı düşünülmektedir. Ayrıca kitapta yer alan örnek uygulama ve sorularda gerçek veri setleri ve güncel örnekler kullanılmıştır. Örneklerin tamamında uygulama adımları en baştan itibaren verilerek uygulamaların anlaşılabilirliği arttırılmıştır. Çalışmada Python programlama dili, öğrenmesi ve kullanması kolay olması yanında diğer programlama dillerinden daha az ayrıntılı ve okunaklı olduğu için tercih edilmiştir. Özellikle Python programında uygulamalar sonrası elde edilen çıktılar, analistler ve karar vericiler için tatmin edici olmaktadır.

İÇİNDEKİLER

1.BÖLÜM

PYTHON KURULUMU ve KÜTÜPHANELER 1

Python Kurulumu	1
Python İstatistik Kütüphanelerinin Seçimi.....	8

2.BÖLÜM

PYTHON'DA VERİ TÜRLERİ VE VERİ GİRİŞİ 11

Veri Türleri ve Veri Girişi	11
Kategorik Veri Türleri.....	11
Sayısal Veri Türleri	11
Python' da Veri Girişi	12
Tuple (Demet) Veriler	12
List (Liste) Veriler	12
Array (Dizi) Veriler.....	13
Dictionary (Sözlük) Veriler	13
DataFrame (Veri Şablonu) Veriler	14
Series (Seri) Veriler	17
Rastgele Sayı Üretimi	19
Python'da Dosya ile Veri Girişi	21

3.BÖLÜM

PYTHON'DA KULLANILAN TEMEL İFADELER VE DÖNGÜLER 29

Python'da Kullanılan Temel İfadeler ve Döngüler.....	29
“if” İfadesi	29
“for” Döngüsü.....	35
“range()” Döngüsü	37
“for” Döngüsünde “else” İfadesi	39
“while” Döngüsü	40
“def” İfadesi.....	42

4.BÖLÜM

PYTHON'DA VERİLERİN İŞLENMESİ 47

Verilerin İşlenmesi	47
Hiyerarşik İndeksleme.....	47
Veri Kümelerini Birleştirme.....	55



Veri Şablonu (DataFrame) Birleştirme.....	56
Dizinleri Birleştirme.....	63
Bir Eksen Boyunca Birleştirme.....	68
Verileri Örtüşme ile Birleştirme.....	74
Verileri Yeniden Şekillendirme ve Döndürme.....	77
5.BÖLÜM	
VERİLERİN TEMİZLENMESİ VE HAZIRLANMASI.....	83
Verinin Temizlenmesi ve Hazırlanması.....	83
Eksik Verilerin Düzenlenmesi.....	83
Yinelemelerin Kaldırılması.....	88
Verilerin Değiştirilmesi.....	89
Satır ve Sütun İsimlerinin Değiştirilmesi.....	91
Kukla Değişkenler.....	92
Verinin Dosya Çıktısı Olarak Alınması.....	94
6.BÖLÜM	
TANIMLAYICI İSTATİSTİKLERİN HESAPLANMASI.....	95
Popülasyon ve Örneklem.....	95
Örnekleme Yöntemleri.....	97
Olasılığı Bilinmeyen Örnekler.....	97
Olasılık Örnekleri.....	97
Merkezi Eğilim Ölçüleri.....	98
Aritmetik Ortalama.....	98
Ağırlıklı Ortalama.....	100
Harmonik Ortalama.....	102
Geometrik Ortalama.....	103
Mod.....	104
Medyan.....	107
İki Boyutlu Verilerle Çalışma.....	109
Değişkenlik Ölçüleri.....	113
Varyans.....	113
Standart Sapma.....	115
Çarpıklık.....	116
Basıklık.....	118
Yüzdelikler.....	120
Aralıklar.....	122
Tanımlayıcı İstatistiklerin Özeti.....	123

7.BÖLÜM**VERİLERİN GÖRSELLEŞTİRİLMESİ 127**

Veri Görselleştirme	127
Çizgi Grafikleri.....	128
Birikimli Alan Grafikleri.....	131
Kutu Grafikleri.....	132
Histogramlar	137
Lolipop Grafiği	144
Pasta Grafikleri.....	145
Çubuk Grafikler.....	148
Dağılım Grafikleri.....	152
Isı Haritaları.....	156
Üç Boyutlu Grafikler.....	160

8.BÖLÜM**OLASILIK DAĞILIMLARI..... 167**

Olasılık Dağılımları.....	167
Bernoulli Dağılımı	167
Binom Dağılımı	169
Poisson Dağılımı	170
Normal Dağılım	171
Görsel Normallik Kontrolleri	173
Histogram Grafiği.....	173
Q-Q Grafiği	174
İstatistiksel Normallik Testleri	175
Shapiro-Wilk Testi	176
D'Agostino'nun K^2 Testi	176
Anderson-Darling Testi	177

9.BÖLÜM**AYKIRI DEĞERLERİN TESPİTİ VE KALDIRILMASI..... 179**

Aykırı Değerlerin Tespiti	179
Görselleştirme.....	179
Kutu Grafiğini Kullanma	179
Dağılım Grafiği Kullanma	181
Z puanı Kullanma	182
Çeyrekler Arası Aralık (IQR) Kullanımı.....	183
Aykırı Değerlerin Kaldırılması	185

10.BÖLÜM**HİPOTEZ TESTLERİ..... 187**

Hipotez Testi	187
Sıfır Hipotezi ve Alternatif Hipotez	188
Hipotez Testinde Kullanılan Temel Kavramlar	188
Güç Analizi	191

11.BÖLÜM**PARAMETRİK HİPOTEZ TESTLERİ..... 197**

Parametrik Testler	197
Z Testi	197
Tek Örneklem Z Testi.....	197
Tek Anakütle Ortalamasının Z Testi	197
Tek Anakütle Oranının Z Testi	202
İki Örneklem Z Testi	205
İki Anakütle Ortalamasının Z Testi.....	205
İki Anakütle Oranının Z Testi.....	208
Student T Testi.....	212
Tek Örneklem T-Testi.....	213
İki Örneklem T-Testi.....	216
Örneklerin Bağımsız Olma Hali.....	217
Eşlenik - Çift Örnekler Hali	220
Varyans Analizi (ANOVA) (F Testi).....	224
Varyans Homojenlik Testleri.....	226
Bartlett Homojenlik Testi	226
Levene Homojenlik Testi	227
Çoklu Karşılaştırma Testleri.....	227
Parametrik Karşılaştırma Testleri.....	227
Tukey Testi.....	228
Scheffe Testi	228
Student T Testi	228
Tamhane T2 Testi.....	228
Tek Yönlü ANOVA.....	228
İki Yönlü ANOVA	238
Tek Yönlü MANOVA.....	249
İki Yönlü MANOVA	251

12.BÖLÜM**PARAMETRİK OLMAYAN (NONPARAMETRİK) HİPOTEZ TESTLERİ ... 255**

Nonparametrik Testler.....	255
Tek Örnekli Verilerin Analizi	255

Ki- Kare Testi.....	255
Tek Örneklem Ki-Kare Testi.....	257
Bağımsızlık Testi	259
Yates Düzeltmesi	259
Fisher'ın Kesin Ki-kare Testi	261
McNemar Testi.....	262
Kolmogorov-Smirnov.....	264
Bağımsız İki Örneğin Verilerinin Analizi.....	267
Mann-Whitney U Testi	267
Kolmogorov-Smirnov Testi	272
Wald-Wolfowitz Dizi Sayıları	277
İkiden Fazla Bağımsız Örneğin Analizi	278
Parametrik Olmayan Karşılaştırma Testleri	278
Conover Testi	279
Dunn Testi	279
Kruskal Wallis H Testi.....	279
Mood Medyan Testi.....	283
İlişkili(Eşlenik-Çift) İki Örneğin Analizi.....	287
Wilcoxon İşaretli Sıra (Eşlenik) Testi	287
İkiden Fazla İlişkili Örneklerin Analizi.....	291
Parametrik Olmayan Bağımlı Grupların Karşılaştırma Testleri.....	291
Nemenyi Testi.....	292
Conover Testi	292
Friedman Testi.....	292
Cochran Q testi	295

13.BÖLÜM

REGRESYON VE KORELASYON 297

Regresyon ve Korelasyon.....	297
Veri Çiftleri Arasındaki İlişki Ölçüleri.....	297
Korelasyon	297
Pearson Korelasyon	299
Spearman Korelasyon.....	302
Kendall Tau Korelasyon	303
Kovaryans.....	304
Regresyon Modelleri.....	305
Heteroskedastisite Sorunu	310
Basit Doğrusal Regresyon.....	311
Çok Değişkenli Regresyon.....	321
Polinomal Regresyon.....	334
Lojistik Regresyon.....	348

14.BÖLÜM	
YAŞAM (SAĞKALIM) ANALİZİ	351
Yaşam Analizi.....	351
Kaplan-Meier Sağkalım Analizi	352
Nelson-Aalen Tehlike (Hazard) Oranları Tahmini.....	370
Log-Rank Testi.....	375
Cox Orantılı Tehlike Modeli.....	377
15.BÖLÜMFAKTÖR ANALİZİ.....	385
Faktör Analizi	385
16.BÖLÜM	
KÜMELEME ANALİZİ	391
Hiyerarşik Kümeleme	391
Hiyerarşik Olmayan Kümeleme	397
K-Ortalamalar (K-Means) Yöntemi	397
KAYNAKLAR.....	401

KAYNAKLAR

- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901), 268-282.
- Bortz, J. (1993). Statistik für Sozialwissenschaftler. 4. Aufl., 419-420.
- Conover, W. J., & Iman, R. L. (1979). *On multiple-comparisons procedures* (pp. pp1-14). Technical report, Los Alamos Scientific Laboratory.
- Cressie, N., & Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3), 440-464.
- D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2), 341-348.
- Downey, A. (2014). *Think stats: exploratory data analysis*. "O'Reilly Media, Inc."
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241-252.
- Glantz, S. A. (2002). Primer of biostatistics.
- Haslwanter, T. (2016). An Introduction to Statistics with Python. *With applications in the life sciences*. Switzerland: Springer International Publishing.
- <https://www.kdnuggets.com/2020/07/complete-guide-survival-analysis-python-part1.html>
- <https://www.kdnuggets.com/2020/07/guide-survival-analysis-python-part-2.html>
- <https://www.kdnuggets.com/2020/07/guide-survival-analysis-python-part-3.html>
- <https://www.kaggle.com/kashnitsky/topic-9-part-1-time-series-analysis-in-python>
- <https://numpy.org/doc/stable/reference/routines.statistics.html>
- <https://docs.scipy.org/doc/scipy/reference/stats.html>
- <https://pandas.pydata.org/pandas-docs/stable/index.html>
- <https://matplotlib.org/stable/users/index.html>
- <https://pypi.org/>
- <https://www.anaconda.com/products/individual#Downloads>
- Idris, I. (2014). *Python data analysis*. Packt Publishing Ltd.
- Kaplan, D. (2009). *Statistical modeling: A fresh approach*. St Paul: Macalester College.
- Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS one*, 10(8), e0132382.
- Kendall, M. G. (1948). Rank correlation methods.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583-621.
- Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.
- Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., ... & Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12(3), 601-607.
- Lothar S., (1997). *Angewandte Statistik*. Berlin: Springer. Pages: 395-397, 662-664.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
- McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. "O'Reilly Media, Inc."

- Mukhiya S. K., Ahmed U. (2020). *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd.
- Pratt, J. W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287), 655-667.
- Rajagopalan G. (2021). *A Python Data Analyst's Toolkit: Learn Python and Python-based Libraries with Applications in Data Analysis and Statistics*. Switzerland: Springer International Publishing.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1), 21-33.
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2), 87-110.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611.
- Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics*, 357-369.
- Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, 74(366a), 471-480.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99- 114.